<Bioinformatics Center>Bio-knowledge Engineering

Bioinformatics Center
– Bio-knowledge Engineering –

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
KARASUYAMA, Masayuki
(D Eng)

Assist Prof
NGUYEN, Hao Canh
(Ph D)

Proj Res
NATSUME, Yayoi
(D Agr)

Students

TAKAHASHI, Keiichiro (D3)
MOHAMED, Ahmed (D2)
YOTSUKURA, Sohiya (D1)
YAMAGUCHI, Shigeru (D1)
CHEN, Zhuoxin (M2)

Visiting Researchers

Lecturer HIBISHY, Hanaa	Tanta University, Egypt, 17 February–15 August
Dr. KUSTRAZE, Ia	Ilia State University, Georgia, 8 April–5 July
Mr. JOHNSTON, Ian	Boston University, U.S.A., 30 May–19 August
Mr. YU-DE Chen	National Cheng Kung University, Taiwan, 17 September–17 March

Scope of Research

We are interested in graphs and networks in biology, chemistry and medical sciences, which include metabolic networks, protein-protein interactions and chemical compounds. We have developed original techniques in machine learning and data mining for analyzing these graphs and networks, occasionally combining with table-format datasets, such as gene expression and chemical properties. We have applied the developed techniques to real data to demonstrate the performance of the methods and further to find new scientific insights.

KEYWORDS

Bioinformatics
Computational Genomics
Data Mining
Machine Learning
Systems Biology

Selected Publications

Zheng, X.; Ding, H.; Mamitsuka, H.; Zhu, S., Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions, *Proceedings of the Nineteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013)*, 1025-1033 (2013).

Nakamura, A.; Saito, T.; Takigawa, I.; Kudo, M.; Mamitsuka, H., Fast Algorithms for Finding a Minimum Repetition Representation of Strings and Trees, *Discrete Applied Mathematics*, **161 (10-11),** 1556-1575 (2013).

Takigawa, I.; Mamitsuka, H., Graph Mining: Procedure, Application to Drug Discovery and Recent Advance, *Drug Discovery Today*, **18 (1-2),** 50-57 (2013).

Mamitsuka, H.; DeLisi, C.; Kanehisa, M., Data Mining for Systems Biology: Methods and Protocols *Methods in Molecular Biology*, **939,** (2013).

Hancock, T.; Takigawa, I.; Mamitsuka, H., Identifying Pathways of Co-ordinated Gene Expression Data Mining for Systems Biology: Methods and Protocols, *Methods in Molecular Biology*, **939,** 7, 69-85 (2013).

Global Graph Comparison for Biological Networks

We investigate the new problem of global graph comparison from statistical viewpoint, with the application in studying evolutions and preservations of biological networks of different species. Previous works on comparing graphs mainly focused on comparing graphs locally, considering graphs as a collection of subgraphs (as in the case of large chemical structures consisting of many independently functioning substructures). This does not satisfy the requirement of our application in comparing species through their corresponding metabolic, signaling or protein-protein interactions networks. Instead, in our application, the global structures of networks, such as connectivity and robustness of the networks determine the species' biological functionalities. This can contribute to building phylogenetic trees.

We formulate this problem as a graph comparison problem for labeled graphs. Considering orthologous genes from different species as the same node in different graphs, the problem boils down to comparing different graph structures on the same node set. Taking into account biological interpretation of network connectivity and robustness, we require that graph pairs are similar if they differ in the well-connected parts, and similar in sparse parts. For example, in Figure 1, graph G1 should be close to G2, and far from G3, even though both pairs have one edge difference. Therefore, we use eigenvectors and eigenvalues of graph Laplacians to derive similarities and distances (*ged*) for graphs globally [1,2].

Our formulation for this problem is shown to have many properties. It is shown to have the ability to weight edges in graphs according to their roles in network structures, potentially showing the important steps in biological processes. It is a generalization of comparing embeddings of graphs with graph Laplacians, paving way for more extensions with desirable statistical properties [2]. It has unexpected applications beyond our initial intention. We can also use it to select graph-cut clustering solutions, making it a general tool for graph data analysis. As shown in Figure 2, *ged* can differentiate bad clustering solutions with disconnected clusters from good ones [2]. This cannot be seen from usual clustering algorithms. The next step would be applying the method to comparing biological networks of different species to studies their evolutions and preservations in terms of these biological networks.

References

[1] Wicker, N.; Nguyen, C. H.; Mamitsuka, H., A New Dissimilarity Measure for Comparing Labeled Graphs, *Linear Algebra and its Applications*, **438-5,** 2331–2338 (2013).

[2] Nguyen, C. H.; Wicker, N.; Mamitsuka, H., Selecting Graph Cut Solutions via Global Graph Similarity, *IEEE Transactions on Neural Networks and Learning Systems* (in press).

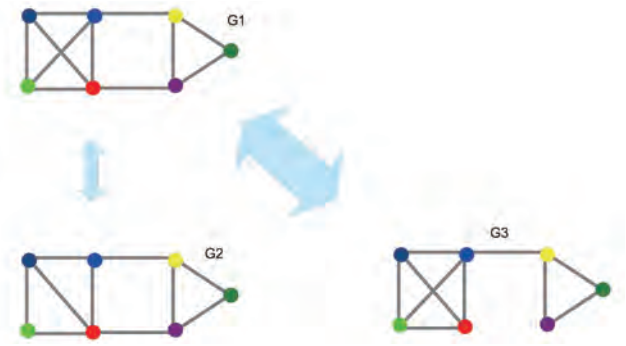

Figure 1. G1 is closer to G2 then to G3 as G3 is not robust, can be disconnected by removing one edge.

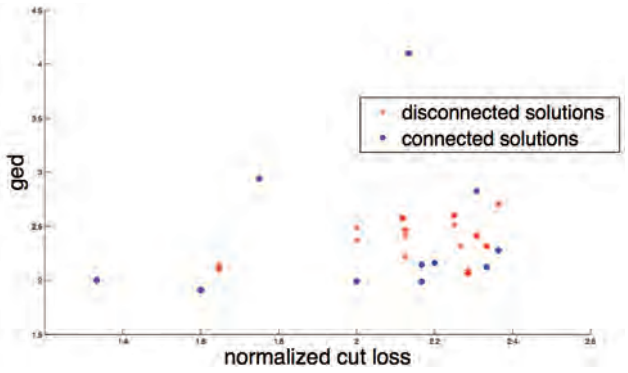

Figure 2. Small *ged* distances usually mean good clustering solutions (connected clusters).